Reusing Implicit Cooperation, A novel approach to knowledge management,

Luigi Lancieri

tripleC 2(1): 28-46, 2004
ISSN 1726-670X
http://tripleC.uti.at

Reusing Implicit Cooperation. A novel approach to knowledge management.

Luigi Lancieri

France Telecom R&D, 42 rue des coutures 14000 Caen, France; luigi.lancieri@francetelecom.com

Abstract: The study described in this paper deals with information reuse obtained by implicit co-operation, particularly by recycling the contents of a proxy cache (shared memory). The objective is to automatically feed a Web server with large multimedia objects implicitly centred on community fields of interests. We show that the strategy of reusing previously downloaded information provides interesting advantages at a low cost; in particular, to reduce Web access time, to improve information retrieval, and to reduce Internet bandwidth use. Moreover, we use the conceptual frameworks of forgetting and collective intelligence to develop a model on which the operation of implicit cooperation is based.

Keywords: Human Factors, Human-computer interaction, community behaviour, Collaborative computing, Shared memory, Data Mining;

Acknowledgement: Thanks to my colleagues N. Durand, S. Legouix, N. Saillard P. Mansson , S. Lenoir and to the TripleC reviewers for theirs help and constructive comments.

1 Introduction.

From time immemorial, humans have sought to extend the capability of their memory, whether individually or collectively. From this point of view, the era of printing was fundamental, but not only because books are a huge and reliable physical form of long-term memory (a shield against forgetting). It seems obvious that the spread of each media, each one in its own way and at each age of humanity, has had a high impact on social change. In fact, this development follows a self-referential logic: new media result in social change that, in turn, creates media evolution that, in turn, promotes social change, and so on. This dialectical process can have as thinkers like e.g. Vilém Flusser have shown depending on the form of human intervention had positive and/or negative social consequences. Books and computers today are material extensions of human memory that allow us our knowledge to be shared to a larger extent than through direct human interactions. This shared memory takes an important place in the phenomenon of collective intelligence.

This situation is taking on a new dimension nowadays with the spread of interconnected computers. Everyone can experience the fact that the emergence of the Web offers new opportunities in our relationship to information and our interactions. Pierre Levy (1994) says:

"..knowledge could once again be carried by living human collectivities rather than by separate material bases. With the difference that this time, the immediate carrier of knowledge would no longer be the physical community with its carnal memory but cyberspace, the region of virtual worlds through the intermediation of which this community would recognize its objects, and itself as a collective intelligence." (See reference section for details). Internet, indeed, showed its capacity to make the most of individual intelligence by collective action (news groups, free software initiatives, etc.). But, even if collective intelligence seems to be a powerful mechanism, it appears to be, mainly, latent, uncontrollable, and consequently difficult to exploit. Besides the high availability (a lot of information, search engines, etc.) of shared information, interconnected computers also allow easy reusability. This concept is not really a new one, but it is becoming more applicable in the electronic information era. The idea of reusability is obvious, considering that (compared to books) electronic documents can be easily duplicated and diffused. Less obvious is the concept that we will develop of reusing collective intelligence imbedded in electronic information.

The goal of this paper is to describe a practical example of this kind of reusability. As industries take advantage of existing natural energy (wind, sun, etc.), we will see how collective intelligence and cooperation can be reused in order to solve real and difficult technical problems. The problems we will examine relate to time wasted in accessing information (searching time and downloading time). Paradoxically, this loss of efficiency results from the success of the Web. Due to an enormous quantity of data and to the increase of users, the Web has become slow and it has become difficult or takes a long time to find pertinent information. As we will see in the state of the art, there is no unique technical solution today that settles these two problems, which are usually approached separately. In this paper we will show, through a multidisciplinary approach, how the two problems are linked and how it is possible to approach them using a unique and basic strategy taking advantage of collective intelligence. To be as clear as possible, it will be necessary to give some technical explanations about a regular network device named “Proxy cache”. We will take advantages of some little-known characteristics of this device in order to capture, catalyze and reuse user cooperation. We will reuse the contents of such a device that result from prior downloads by a subset of a community of users. We will see why we can say that reusing these contents can be seen as collective intelligence reusability and how this strategy can solve our problems.

Some experts think that the access speed to information could be improved by a suitable dimensioning of the network (access, backbone, etc.). Our opinion is that this approach is not sufficient because the network latencies are not solely explained by the weakness of the links. For example, even if a user is on a good-sized network, his download will be slow if the object comes from an overloaded server or through a saturated distant network. We think that the duplication of large and pertinent objects in proximity to the end users is a much more effective approach. Some techniques such as proxy caches, Web mirror servers or CDN (Content Delivery Network) based on this principle, associated with fast access networks (xdsl, cable, etc.) allow us to surf the Web with an improved response time. However, despite their advantages, the underlying problem of these techniques is that their efficiency depends on the fact that the desired object needs to be identified and stored before the user asks for it. This may appear evident, but, even if replicating information on servers close to the users seems to be a good solution, the problem is to know and to anticipate what information is useful to replicate and how to feed these local servers so that the economic output remains favourable. The problem with information selectivity is usually solved by search engines, but these devices index only a small part of the Web (about 10 %, see Lawrence et al (1999)) and show a lack of pertinence in their results. In spite of their progress, most of the search engines return a lot of irrelevant results and

do a poor job indexing multimedia contents. Furthermore, engines return only references and the downloading time of the corresponding distant large objects can remain very long.

In our study, the users of a local area network (LAN) are connected to the Internet through a shared proxy cache. As a mediation device, the proxy cache relays the LAN users' requests and stores the resulting objects locally, anticipating, in terms of probability, requests of other users, all of which consequently speed up the access when the user requests refer to cached pages. On average, when a user makes a request to a distant web server, he/she has a probability of about 30 % (Davison (1999), Rochat (1999), see also Web Replication resources) that the object is really in the cache. In fact, the more the users share common interests, the higher this probability and the resulting access speed are. In other words, the access speed improvement is a direct result of the implicit cooperation level between users. One of the main problems, from our point of view, is that the users do not know what is stored in the cache (contrary to a Web server) and so it is difficult to explicitly reuse cached objects. Thus, according to what we will see in the state of the art, proxy caches are not usually used to help explicit information retrieval, but only as probabilistic access accelerators.

Furthermore, according to its role as a mediator device, we can observe that the proxy cache keeps the context of the download and the most popular objects over a period of time. We postulate that these popular objects, which represent a kind of consensus, could possibly interest other users of the LAN community. We will show that the validity of this assumption highly depends on the behaviour of the community and we will introduce a measure to evaluate the efficiency of such an approach. Therefore, the main function of our system is to regularly collect the cache's largest and reusable multimedia objects, in order to put them in a local web server with a user-friendly interface. We will show that, as a "human-system" interaction, the combined effect of user download activities through the limited shared memory of the cache (need of "forgetting" effect), produces what is called a symbiotic mechanism (see Section 2) that catalyzes the collective intelligence of the group of users. One of our contributions is to identify this latent "natural" mechanism and to propose means of reusing it in a wider context.

This paper will start with a theoretical discussion (section 2) that will describe the concepts underlying our approach. Section 3 will describe the proxy cache mechanism and show why it is an essential device as a human-system interaction captor. Sections 4 and 5 will describe the architecture of our system and the first results we obtained in an experimental context. We will also see in Section 6, that reusing objects that have previously been downloaded can raise ethical and legal problems, and point out some possible solutions. Finally, before concluding this study, we will survey a utilization of implicit cooperation in the Web (Section 7).

2 The Conceptual Approach

Edgar Morin (Cabin (1998)) says that knowledge contextualizes information. He says that as self-adaptable filters, a "thinking being" (through action and interaction) creates knowledge as a reduced set of a lesser degree of contextualized knowledge (called here information but that could be real knowledge in a better suited context). Knowledge creation is, at first, an operation of information space reduction. In this definition, knowledge can emerge from cooperation between humans. The notion of cooperation can have different meanings and is used in a lot of contexts (See C. Fuchs (2003) for a state of the art and a discussion). This notion can involve varying degrees of consciousness, autonomy or spontaneity from individuals with more or less predictable results. Contrary to explicit cooperation where people collaborate openly (e.g. several writers for a single document), implicit cooperation involves the fact that people cannot easily evaluate qualitatively and quantitatively the contribution of others. It would be too strict a definition to qualify implicit cooperation as dialectic, but it is a good approach.

We think that there are both implicit and explicit components of cooperation within complex human systems. However, the part of each of those components is different on almost each level of this system, ranging from the physical or biological to the social or societal. So that it is not possible to comprehend human cooperation without taking into account both aspects. Following Herbert Simon (1983), human actions are not always rational; some influences are sometimes unconscious or biased by a necessarily approximate perception of reality (i.e., limit of the cognitive system) making the result of cooperation a complex mixture of explicit and implicit involvement. The Art practice (creation and understanding) seems to be a good example of such complexity. Also, Dan Sperber (1996) describes how ideas, rumours or culture are going through people as along a chain, gradually reproducing themselves locally. Some other theories, such as social Darwinism or socio-biology (Edward O. Wilson), also see the social operating mode as a complex association or fusion of very simple mechanism (see also G. Edelman (1992)). From our point of view, all these basic components (e.g. forgetting), as in iterative systems, are similarly causes and consequences of what could be called collective intelligence.

Cooperation and collective intelligence are well known in the living world and in particular in human societies. The basic idea is that cooperation between individuals is more productive than the sum of individual actions. We think that there is a close link between collective intelligence and implicit cooperation combined with forgetting effects.

Pierre Levy pointed out that the concept of collective intelligence should remain in the context of human activity excluding, for example, mechanisms of cooperation in the animal world. I partially agree with this position. Like him, I think that human societies remain the ultimate context where collective intelligence can be the most "effective" in all domains of life. Furthermore, all efforts should be made to preserve and enhance the human individual as well as the collective. But considering the question on how collective intelligence works, I think that the observation of human contexts is too limited a solution because the problem is too complex. A side approach on artificial systems or animal beings is very instructive on all aspects involved in human cooperation, even those that are mechanical or not easily observed.

We will now give several examples, from the simplest to more complex, involving implicit cooperation and showing the role of the forgetting effect. The first example, based on animal cooperation, is not trying to say that animal cooperation is the same as the human one. In my opinion, "animal like" cooperation is simply a cooperation component of a more complex cooperation system operating in human societies.

Let us take the example of ant colonies that reach a high level of organization with a low level of individual “intelligence” thanks to collective action. What is very surprising with ants is their capacity to rapidly find the shortest path to food. Studies (Bonabeau (2000)) show that the “secret” reason for this capacity is, paradoxically, the loss of information that we call the “forgetting effect”. Let us explain this phenomenon. At the beginning, we have a colony of ants looking for new sources of food that cannot be pinpointed. Therefore, we observe what seems to be a random individual trip of ants between the anthill and the food that rapidly becomes a unique straight line for all the individuals of the colony. We have known for a long time that ants leave pheromone on their paths. Following the information linked to the persistence of this odorous substance helps them to go back to their anthill. Since the pheromone evaporates (loss of information), the shortest path odor lasts longer than the others (shortest reinforcement delay). Thus, over time the ants have a higher probability of using the shortest path that contributes to its reinforcement. It is important to understand that the high organizational capacity of the entire group is mainly possible thanks to the basic property of the forgetting effect. This example also illustrates the symbiotic effect (De Rosnay (1995)) that involves interaction between "intelligent" actors (human, animals) and material (ground and pheromone, computers, etc). This kind of mechanism is also used in “brain storming” meetings that are known to be a very efficient use of collective intelligence to solve problems. Keeping the essential (consensus of the group) elements on the blackboard and deleting

progressively the other pieces of information, also involves a kind of forgetting effect. It is interesting to note that the sharing of ideas or simple informal discussion are not sufficient to find the solution. The methodology in brain storming meetings has the explicit goal of helping the group to eliminate (forget) "noise" in order for really useful information to emerge. The Web is a place where informational cooperation, forgetting and symbiotic effects are omnipresent. Some services like news groups, proxy cache or Web hyperlinks are based on more or less implicit cooperation and forgetting (Lancieri (2000)).

Web site cross-references link inter-connecting information from different sources. Hence the Web is the result of a huge cooperation process. From this point of view, we can see the Web as a big informational ecosystem (Wolpert et al (1999)). Although it implies different movements of information with regard to the replication, the link included in a Web page can be viewed as a virtual replication. Indeed, users, depending on personal interest or on the links' popularity, replicate it more or less intentionally. Some of these virtual replicas can be privileged, for example, for intellectual property reasons, due to Web server resource optimization, or depending on content management opportunities. The analysis of the dynamics of these replications underscore the phenomenon of selective accumulation of replicas, corresponding to related subjects and contributing to form complex sets of data reinforcing the same themes or reducing (forgetting, i.e. drown in noise). Through human action, these replicas cooperate in the same direction or they confirm themselves mutually by targeting each other, with more or less strength, according to their semantic proximity or objectives. Two rival commercial documents, for example, will not point to one another. On the other hand, two documents contributing to the defence of the same ideology will make a mutual reference. The more links to a document that can be found, the more people will probably visit it and the more people will refer to it by new links. This mechanism will tend to expand some knowledge and reduce (forget) others. Some authors qualify this cross-referencing phenomenon as self-catalytic (Heylighen (1996)).

Through its consultation, basic information (data consulted) also loads itself with implicit contextual informational imprints that go further than basic content. For example, the knowledge that someone reads mainly adventure books gives larger information than that contained in the books that were read. This new informational dimension linked to the context of downloading (Intentional action) has a high potential regarding individual and collective intelligence. This capacity of extracting from data other information than the one contained in the data itself is an important feature of human cognitive capacity (associative, episodic memory, etc.). D.L Shactker (1996), a specialist on psychological and biological aspects of human memory, says, "Old memories define our personalities". He illustrated that the capacity to take into account the past (contextual episodic dimension of information) means that experience is essential for our cognition. A suitable spatial and temporal management of such an "upper layer" informational set allows the extraction of causality and intentionality contexts in human access to knowledge. In the case of data resulting from cooperation, the potential is higher since the challenge is reusing collective intelligence. Once the advantage of taking part in this upper layer set of information is understood, the entire problem is to catch and manage it in an efficient and inexpensive way.

In the next section, we will see that reusing the content of proxy caches can give us this capability.

3 Reusing Collective Intelligence by Making Use of a Proxy Cache

In this section, we present first the regular features and uses of a proxy cache, then the complementary uses we studied such as a cooperation captor and booster device. Proxy caches are complex devices mainly based on the use of "limited" memory shared between users. A proxy cache combines two functions: the function of relaying the accesses of a group (mediation) and the function of limited shared memory that allows "remembering" web objects close to users. Upon first access to a distant document, this access is memorized locally and returned when other users call it up again, so that data transmission

time and network bandwidth can be saved. The more common fields of interest the users share (implicit synergy) the higher the probability of finding locally a previously downloaded object. In other words, access redundancy is a witness of the level of synergy between users validated by the level of improvement in access speed. This capacity of implicit cooperation capture helps to automatically collect the most popular downloads or the consensus of user activity. The concept of reuse of implicit cooperation and collective intelligence comes from the explicit reuse of proxy cache content (information itself and linked contextual imprinted information) that results from implicit cooperation or user synergy.

Caches have two types of memorization strategies: explicit and implicit. First, a cache stores user accesses as events. The log files contain at least the IP address of the users, the URL of the required object, the date, and the level of success (i.e. a "hit" if the downloaded object was previously requested, a "miss" if not). Every user query generates a line in this log file. This kind of file also exists on the Web servers, but it is less interesting because the accesses relate only to the servers' limited contents. The proxy cache log files relate, on the other hand, to wider open contents (potentially that of the whole Web) and consequently are more expressive of user interests. Another element of much less systematic memorization is that of the Web objects themselves (text, video, audio, etc.). Apart from user behaviour (implicit cooperative memorization), the decision to memorize a document or not, can be taken at the level of the proxy cache or that of the original server for technical or copyright reasons. For example, a page containing personalized data (search engine) or an ephemeral validity (stock exchange) is mostly generated dynamically and is not memorized. As we will see (section 6), these considerations are important as far as copyrights are concerned.

It is also important to see that the "objects'" shared memory of the cache is limited: it is thus necessary to decide, upon saturation, which document to eliminate or to keep. This decision process corresponds to the replacement policy. It is an important aspect of cache performance, which is often subject to a heavy traffic load (to process, in real time, all the queries of the LAN site). The existing literature supplies a lot of information on these algorithms (Kelly et al (1999)) mainly based on stacks as LFU (Least Frequently Used) that has a functioning based on the "forgetting" effect (Lancieri (2000)). Applied to a limited shared memory, this feature allows a consensus to emerge (information mostly shared most of the time). In general, this duration depends on the cache disk size, on the rate of use of the memorized documents, and on the spectral width of the queries (thematic variety). A study carried out by CacheFlow Corporation (See Web resources) shows that in an average case, 66 % of the pages are kept less than 24 hours. As we will see, this relatively short life span is an argument to save potentially interesting large objects on a dedicated server. Let us also say that it is not possible for users to directly browse proxy cache content for several technical reasons. The autonomous active mirror is thus a technical solution to reuse proxy cache content and its correlated imprint collective intelligence, in an efficient and inexpensive way.

4 Description of the Autonomous Active Mirror

Figure 1 shows our system's utilization context. We can see users connecting to the Internet through a proxy cache and a local Web server hosting the management module of our system as well as the large-sized object collected from the cache.

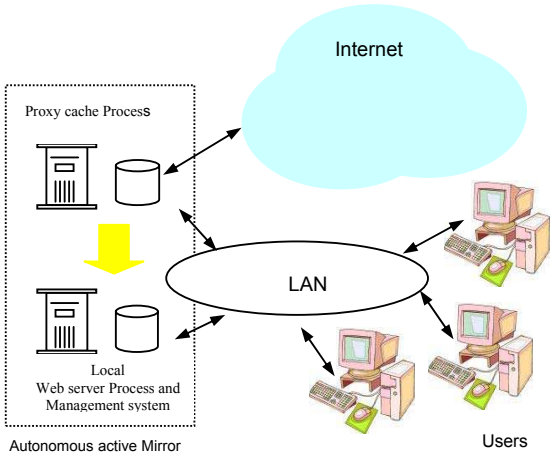

Figure 1: Network architecture

As described in figure 2, the autonomous active mirror contains three main modules. The first one is the proxy cache already in use in the LAN. In addition to its initial function, and in accordance with what we said previously, the proxy cache is used as a captor of knowledge interaction. It needs no special modification except to increase the standard limitation on large files (easy to change in the proxy configuration file). The second module is the system management that we will describe in the next paragraph. The last module is a regular Web server.

The user sees the Autonomous mirror as a regular Web server that he/she reaches via a regular Web browser (see figure 4). The HTML pages located on this server are automatically generated by the management module and contain links to large objects stored in the server itself, as well as their descriptions recycled from the proxy cache. The management module is in charge of four functions.

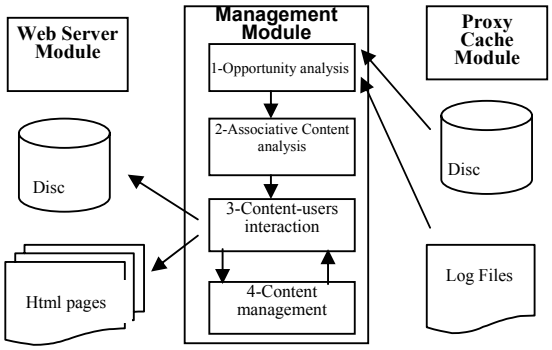

Figure 2: Functional Architecture of the autonomous active Mirror

The first function is to select the part of proxy cache content that will be replicated by the Web server module. At the beginning, these contents (objects, associated HTML pages and log files) are analyzed in order to select only large and easily reusable contents (mp3, mpg, pdf, avi, jpg, etc.) dropping more complex ones (dll, vbx, etc.). Contrary to a regular cache policy, we consider that the cost of storage is lower than the management of replicas. It is thus not worthwhile to recycle small-sized objects since they can be downloaded rather quickly from the original server.

The second function is to describe the recycled objects. Once useful objects have been identified, it is necessary to recopy them and look for some descriptive keywords. This can be carried out in several ways (e.g. URL keywords, analysis of the Web page containing the link to the object, extraction of the informative fields in MP3 files, etc.). The co-occurrence of the most frequent keywords enables us to build a very simple associative word network that allows approximate request (see user interface). This very basic semantic network approach is reasonable here because collaborative filtering obtained through implicit cooperation (user community through the cache) allows us to have a narrow thematic spectrum.

The third function manages explicit interaction with users (to build html pages with description and links to objects, search form, etc.). It is also necessary to build users' profiles that will be useful for managing the contents (mainly life span of objects) and to feed the most interesting objects on the user's personal pages managed by the server module. These profiles are in fact vectors of the most frequent "clean" keywords that come from consulted Web pages or object descriptions. This function also manages authentication access based on an anonymous pseudonym (see part A of figure 4), which is associated with a user profile. A user manual profile description can complement the automated one. Figure 4 shows the Web page as it appears to a user after his authentication. On the right hand side (part D) a search form for keywords and category entry appears (image, music, movie, etc.). It is also possible to see statistics (part F) on the category of recycled objects. This page also contains pictures of the last object recycled consistent with the user profile (part C). It is important to say that the presented objects can be different at each user connection since the flow of collected data from the cache is continuous. Here the server module acts as a regular web server except that the content has been updated and was presented to the users automatically without human intervention. As shown in parts B and E, it is also possible to systematically incorporate general interest information (weather forecast, stock exchange prices, etc.). Each result to the request form is given as a line composed with an icon (video, image, etc) and several descriptive pieces of information (size, most frequent descriptive key words, date) and the link to the local object as well as the link to the original server (copyright purpose).

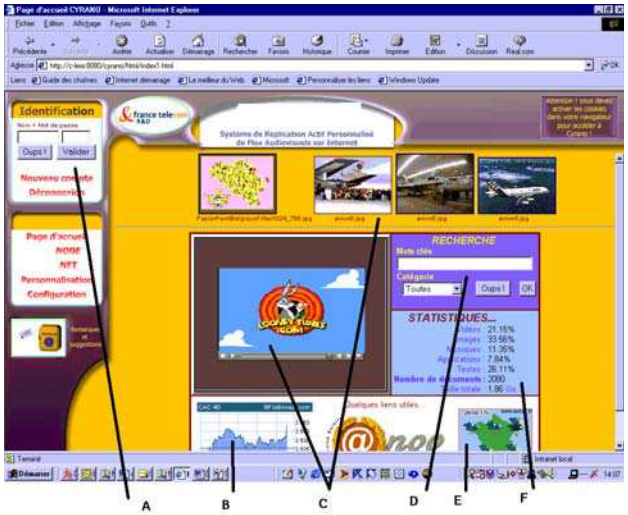

Figure 4: View of the user interface.

The fourth function deals with storage management and the life span of these objects. One of the important differences between our system and a traditional Web server is the automatic management of the objects' lifetime. This is necessary because the object feed process (cache module to server module) is continuous. The server takes into account the characteristics of the object (size, category, thematic content, etc.) and the interactions from the users (number of accesses, variation compared to the topic of interests, etc.). With all these elements the management module computes a user-oriented priority function for each object, which will make it possible to decide which one to keep (multi-criteria LFU stack).

Since the server disc size is very large the object lifespan is at least 1 to 2 weeks long (compared to 24 hours with a standard cache).

5 Experiment and Results

The goal of the experiment is to confirm that the Autonomous active mirror enables us to exploit collective intelligence and to evaluate the effective results of this reusability in order to improve information selectivity and access speed. Our experiment was conducted over one month in the following situation: the proxy cache module was the operational SQUID (see Web resources) cache (3 GB disk) of our intranet, relaying 391 users (researchers and engineers in computer science). Only a limited population of 15 users used the server module (4 GB disk) in order for the experimentation to be manageable (interview of volunteers, checking evolution of usages, etc.). Thus, we had an experimental population (15 users) reusing objects downloaded by an operational population (391 users). As we said earlier, the combination of the server module and the cache module increases the hit rate (reuse rate of local objects) compared to a single proxy cache. Furthermore, a higher hit rate automatically improves the access speed to information. The Hit rate here becomes a combination of the implicit action of the users (access to the cache without the knowledge that it contains the object) and the explicit one (access to the server module with the knowledge of what is downloaded locally).

5.1 Influence on the Access Speed to Information.

The measures and the users' experience show that regular access to the Web remains slow. For example, traffic analysis of the French national academic network Renater (see Web resources) shows that access to a distant Web server is on average 300 times slower than the access to a local one[1]. The following analysis shows that the active mirror significantly increases the objects' reuse rate (compared to a single regular cache) and, consequently, improves the average access speed to the data. We will start by analyzing the behaviour of the cache module (operational access), then that of the server module and the combined action of both (the active mirror).

The cache module over the considered period shows that potentially, the reusable quantity of data is relatively high. Indeed, the operational users carried out requests equivalent to 11 GBytes of data where 79 % were unique requests. As we said, we hope to increase the reuse rate while making these objects visible, thanks to the active mirror. It is also noted that the hit rate of large files is relatively weak, 14 % on average, compared to a rate of 36 % of files overall. Let us remember that this value can be considered as an approximate measure of shared interest (implicit synergy). It evaluates the probability for a new request to be found in the proxy cache (i.e. previously downloaded by other users). This implies that the large objects are poorly re-used in a standard single proxy cache. This is unfortunate because they are the most difficult and longest to obtain. One of the possible reasons explaining this fact is that large objects are proportionally much less frequent than small ones (e.g. 93 % of JPEG images weigh less than 40 KB – i.e. the ergonomic size of a picture - with an average of 11 KB). The other reason is the lack of visibility since these objects are poorly indexed by search engines. In some cases they are only discovered by chance.

After having analyzed a standard situation of reuse (the cache alone), we will now evaluate the benefits of the whole system. We notice that the reuse ratio is higher than that of the cache module. Indeed, we have about 10 more requests per user. This is also a first clue about the interest of the recycled contents.

480 for the server module (7,200 requests / 15 users)
53 for the cache module (20,970 requests /391 users)

[1] One year average accesses to global Internet from France. The average retrieval time of a cached object's first byte varies from 9 to 25 ms in a local cache Intranet against from 1 to 7 sec when the cache retrieves it from the original server.

This is not surprising, taking into account the remark made previously. Indeed, the server module improves access to the large objects resulting from implicit consensus since it presents them explicitly, which cannot be done by the cache alone. Furthermore, since the download is extremely fast, the user hesitates less to consult these objects because /he/she does not have to wait a long time for contents that perhaps are not of much interest. We see that the global reuse rate of the active mirror combines that of the cache module (implicit) and that of the server module (explicit).

These results show that the use of local information (objects already downloaded) is optimized with the active mirror. This implies a reduction of outgoing bandwidth to the Internet and high-speed access to information. After doing a little computation, we can see that if the 7,200 recycled objects equivalent to 6484 MBytes had been downloaded directly from the original server on the web (average 200 KB/sec from distant server VS. 1 MB/sec on a local server on a LAN), the waiting latency for all objects would have been:

9 hours regular access to the web from LAN 1 hour 48 min using the active mirror.

Of course, this average difference would be much higher with fast LAN links and overloaded original servers. (See Lancieri (2001 b) for more detailed results)

5.2 Influence on Information Selectivity

One of the strong advantages of this system is its adaptability. Indeed, the content of the cache moves and follows the user download, so that it constitutes a refined information database centred on the users' main topics of interest. The first version of this system was only able to index html pages (like a search engine but restricted to cache content) but it shows interesting performances. (Lancieri 1997, 2000). We have asked a community, composed of 7 experts in network technology not involved in our project, to make ambiguous requests (the word "Network") and to give an evaluation (scale from 0 to 5) value reflecting the level of interest for the first 20 results provided by our system compared to the HotBot Lycos search engine. The average results are revealing:

3.2 points with the Active Mirror 0.7 points with HotBot.

The file size corresponding to URLs returned by the active mirror were 10 times bigger than those of HotBot (128.2 KB [2] on average compared to 13.4 KB). The topic of the requests was Internet network techniques, which are the field of interests of the testers. In comparison, the first results delivered by HotBot were relatively heterogeneous and strongly directed towards trade and advertising; they also had a very small file-size. The first relevant result provided by HotBot is in the 12th position whereas in the first position our system provided documents with strong technical contents. One will note also the difference in the number of provided results, 300 URLs whereas HotBot provided 50,000; this reflects the strong information space reduction of our system. Actually, these results are not surprising because the contents of the cache follow the main users' interests, deleting less popular objects. The negative point is that the implicit filtering used by this system is only effective when research relates to a topic of interest of the group, outside of this area the results are extremely weak.

It is important to say that the goal here was not to have a complete statistical analysis, which would need to have several different thematic contexts of use needing a special dedicated effort. What we tried to do in this part of the experiment was to have a first confirmation on what common sense suggests:

[2] A 128 kb file represents around 50 A4 textual pages.

staying in a homogeneous group (i.e. implicitly capturing its context) reduces ambiguity. This is what we can see every day when discussing specific topics with familiar people; i.e. we do not need to re-explain jargons and situations well known to them but probably not to a stranger.

5.3 Measuring Co-operation Reuse

Taking into account the fact that the effectiveness of our system depends on the implicit level of cooperation among users, it is significant to be able to evaluate this parameter. In addition, this evaluation will help us to monitor the informational interaction and to forecast the recycle policy potential. The level of implicit cooperation can be evaluated by characterizing categories or clusters among the users, each cluster representing a small community of interest. We propose to compute the yield (*Ycr*) of implicit co-operation as the ratio between the actual (i.e. really done) implicit cooperation (*Ecr*) and the potential (*Pcr*). *Ecr* characterizes the homogeneity of a population according to its real activity (URL regularly consulted on the web). *Pcr* characterizes the homogeneity according to the underlying thematic centres of interest (Keywords embedded in this consulted Urls). As shown in the following formulas *Nu* identifies the number of users, *Ncurl*, the number of clusters based on URL and *Ncwords* based on words. We see that a very heterogeneous group will result in the number of clusters that tend towards the number of users whereas a homogeneous group will tend towards one cluster. Consequently, effective cooperation (*Ecr*) is equal to zero % if all users were to make all different requests (URLs) and is equal to 100 % if there is only one cluster (all users made an identical request)

$$Ycr = \frac{Ecr}{Pcr} \qquad \textit{With} \qquad Ecr = 100.\left(\frac{Nu - Nc_{url}}{Nu - 1}\right) \quad \textit{and} \quad Pcr = 100.\left(\frac{Nu - Nc_{words}}{Nu - 1}\right) \qquad (1)$$

Users sharing the same keywords have a certain probability (potential) of sharing the same URLs but this is not sure. The *Ycr* ratio measures this uncertainty. It is important to see that whereas *Ecr* measures the effectiveness of implicit co-operation *Ycr,* measures the capacity of the group for reaching potential co-operation. In other words, *Ycr* is more like a measure of community behaviour.

In order to experiment this concept in our study, we investigated the flows of 319 (potential users of the system) user's requests over a period of 17 months that represent 151,232 URLs corresponding to 84,750 keywords. The algorithm used to cluster the users is an agglomerative method of hierarchical clustering (HAC) (detailed in Ronkainen 1998). From a distance matrix between users, the method successively creates several clusters by agglomerating the most similar groups. This method makes it possible to obtain groups having a good similarity of intra-clusters (the users of a cluster are close and a good dissimilarity of inter-clusters (two different users of two clusters are dissimilar). We chose to calculate the distance matrix between users with the coefficient cosine (see also Matching/Jaccard/Dice/Cosine, etc.; Frakes et al 1992).

We can observe that the keywords as a more granular description result in better distributed groups than URLs. Indeed, with URLs, we have 261 clusters containing only one user and only 2 clusters with more than 5 users (representing 15 users). With the keywords, we have only 39 clusters with only one user and 15 clusters with more than 5 users (representing 192 users). The total number of clusters is 91 for words (*NcWords*) and 280 for URLs (*NcUrls*). Applying the formulas (1) to the results, we obtain the following values:

Effective cooperation:	Ecr=100.(319-280)/(319-1)=12.26 %
Potential cooperation:	Pcr=100.(319-91)/(319-1)=71.69 %
Yield of the cooperation:	Ycr=100.12.26/71.69=17.1 %

Contrary to what one may think, the ideal situation is not a high level of effective cooperation implying that all users have already accessed (i.e. know) the same URLs, which limits the interest in this system. The situation where the potential cooperation is low is not a good case either since it implies that there is a poor probability that the object downloaded by some users could interest others. The best situation is a low effective cooperation and a high potential one. This implies that the system brings new and interesting information to a large audience, which is the case here.

6 Legal and Ethical Aspects

Collecting and reconditioning objects downloaded from the Web poses several legal and ethical problems with respect to users (e.g. maintain access confidentiality) and to the owner or the supplier of the objects (intellectual property). These questions are important and need particular attention. Regarding the ethical and human rights aspects, technologies based on collecting and analysing Internet user activity (browsing paths, analyse of consulted documents, profile, etc.) can be used as a weapon or an instrument of power in malicious hands.

Mainly, the first remark we can make is that even if Internet could be seen as a means of progress, it unfortunately contains, natively, all facilities to go against privacy. As in most cases, technological improvements allow not only human progress but have also sometimes proportionally harmful drawbacks (i.e. cars are useful but kill more people than wars).

All internet-based techniques incorporate monitoring and logging processes. From a personal computer (cookies, browser history or cache, spy ware, etc) to network components (routers, mail, http server or gateway, proxy cache, etc), all is done to allow user activity tracking. Initially, this was done only to guarantee the correct operation of the technology but, with the emergence of new services, some misuse is possible. In fact, this concern is not really new for telecommunication technologies. Since its origin, in the early part of the last century, newly hired post and telegraph employees had to take an official oath, swearing to not reveal the content of private phone calls or telegraphs. It is the same nowadays. For example in a company, only authorized people have access to proxy cache traces that can provide some of the most exploitable information concerning people's activities.

In our systems, there are potentially two sources that possibly could provide information on user activity. The first one is the proxy cache that traces user activity. This is one of the basic functions of a proxy cache that falls under the responsibility of the network administrator. The second source is that of the active mirror modules, even if it is also supposed to fall under the responsibility of the administrator, all users are protected through an anonymous pseudonym. Furthermore, it is important to say that the active mirror does not add privacy intrusion (quite the contrary) from the network’s regular devices.

For ethical reasons, it is always essential that the users be well informed and that they validate the use that is made with their browsing traces. We choose to maintain user confidentiality by using a pseudonym, making traces anonymous. But, whatever the precaution, privacy is not easy to manage because, technologically speaking, it is not possible to totally eliminate all risks of illegal use or misuse of information. Even if every effort should be made, the question remains a philosophical one: does the hope of progress justify the risk of failure?

Beyond privacy, another problem to consider is intellectual property rights and access to unsuited content (child protection). Basically, we may say that the answers strongly depend on the usage context. Indeed, the case of Intranet use with a multi-site International corporation contents is different from contents coming from the Internet and uncontrolled by a public ISP (See on the EC web site the "Data Protection Directive" on related laws for more details)

The first case is rather basic and can be regulated by a company charter knowing that from a legal point of view the problem is not more complicated than that of regular access to the Web. These corporate networks are sometimes very large with such a quantity of information that one finds the same problems as on the Web even if the contents are better controlled. In addition, in most countries legislation allows companies to analyze the traces of connections if the employees are well informed. Thus, the implementation of this system makes it possible to benefit from the company staff's implicit collaboration on “clean” contents. In the case of open Internet access, it should be pointed out that the contents of the system are not different from those downloaded by the employees and stored on their computer. From this point of view, our system does not really modify the responsibility of the company. However, in some cases, the law forces the companies to comply with precise rules that imply, for example, the use of filtering systems to eliminate downloading of undesired contents. The case of the general public I.S.P is more delicate.

Another problem is that we need to satisfy the content suppliers who are often paid according to the number of user hits, which means that the object must be consulted on the original server. The system described is not incompatible with this constraint. As said above, not all objects are cached. In fact, the content suppliers can avoid the cachability of an object. So a content owner or supplier can freely decide (dedicated HTTP option header) if his/her contents can or cannot be reused on other servers. In our system, for example, we computed the cachability rate showing that about 50 % of the large objects are reusable. If the content suppliers enable the reuse function, the active mirror allows us to automatically send back user hits (without object transfer over the network). This method is rather common (see Network appliance and Akamai Web Site) in replication systems. In addition, the URL of the original server is shown to users with the link to the local object.

Thus, the Autonomous mirror acts as an archive collector with all needed references to the content owner. Thus, one can say that a certain number of measures can be taken to satisfy the content supplier as well as the final user.

7 Survey of Implicit Co-operation on the Web

Collective Intelligence and cooperation was the subject of a lot of interesting studies over the last decade (see Web Resources). Briefly we can split these studies into 3 mains areas: the first one stems from sociology and the humanities (Cabin 1998), the second one from artificial system studies (Wolpert 1999) and the last one is a more equilibrate mixture of the first two and basically involves Human-Systems interactions. This last field is the main focus of our studies.

7.1 Cooperation for Increasing Access Speed

Mirroring is without a doubt the most common method in the field of content replication. A Mirror is based on Web server technology and duplicates contents from a main server towards other ones distributed over a given geographic zone. Contrary to the proxy caches, the mirror content management is completely deterministic, i.e. the administrator must take the initiative and explicitly store information and consequently control all parameters related to the contents of these servers (lifespan, quantity of replicates, localization, etc.) (See Web resources). Therefore, its management is often mechanical (in some cases manual) and generally implies identical organizations of the contents between the original server and the mirrors.

Numerous academic studies based on replication aim at optimizing Internet access speed. We will describe here only the initiatives that insist on implicit cooperation reuse and on the phenomenon of collective intelligence, but there are also a lot of very interesting projects and papers from a more quantitative point of view (Danzig 1995, Neal 1996). Many of these studies start from the principle to pre-

fetch the download of objects on caches located near the end users (analyze past downloads to forecast future ones). According to A Bestravos (1996), access anticipation enables the load of the original servers to be reduced by 30%, the user latency period to be reduced by 18% and the hit rate to be improved by 18%. The drawback is a 5 % increase in the used bandwidth (mainly due to forecasting imprecision).

Others studies take into account the geographic situation of the user, the network availability, as well as the user navigation profile (access pattern). (See J. Gwertzman (1995), Tewari et all (1998), LSAM project (1999), or recent (CDN) techniques).

For example, P. Rochat et al (1999) suggest taking into account the cultural influence of the users to optimize performances of caches. The basic postulate is that there is a strong link between the words contained in the URL and the users' centres of interests. The connectivity analysis of these various words enables a neural Kohonen map to be generated, where connections are weighted according to word occurrence frequency in the consulted URL. This method allows documents, semantically similar and frequently accessed, not only to be grouped but also it can be used to optimize the cache content.

In all these studies, knowledge about human interests or behaviour is used to speed up access, but used devices are considered to be as transparent as caches by the end-user (contrary to what happens with the active mirror where the recycled objects appear explicitly). There is no explicit information delivery from these devices. The user surfs the web as he usually does, the difference is that the implicit cooperation allows high probability that an object can be located near the user when he requests it. The user thinks that the delivered data comes from the distant server and he has no means of knowing that the information, in fact comes from a local device.

7.2 Cooperation for Improving Information Selectivity

Contrary to what was developed in the previous section, the following methods allow information to be explicitly delivered to the user but here the objects are not local. Here, the user knows that there is information base and searches it in order to find useful data. Generally speaking, the information selectivity faces two mains gaps in the link between the human and the information. First, it is difficult to know and especially to formalize what exactly a user is thinking when he is seeking information. The second problem is that the huge quantity of available information is difficult to characterize. Linguistics shows us that many ambiguities stem from the fact that a single word (symbol) can refer to many different realities. A lot of academic work tackles these two problems (see for example Hu (2001), Tang 2001, W3C collaboration Research Team 2003).

According to ARIANE (Himmelsbach(1999)), there are about 1,350 search-engines (Google, Altavista, Northern Light, etc.) in the world today. The studies show that the majority of the engines provide between 15 and 42% of documents considered to be relevant (Zaine (1999)). These engines can be classified into five categories mainly differing by content analysis techniques used to index documents (statistical, semantic, natural language, etc.). Relatively few engines take into account cooperative interaction with users. At best, one engine carries out a fine analysis of the words of the request or another one asks the user to specify his request (Arisem, Albert, Alogic, Inktomi search Web site). The most effective approach seems to be collaborative filtering which implies a follow-up interaction concerning accesses and transparent to the user. Furthermore, multimedia-indexing techniques improve their efficiency but remain rather hard to use since they are time- and resource-consuming, thus these techniques are poorly used in search engines.

Exploiting collective intelligence from the Web can also be done by user-oriented re-aggregation from part of distributed Web pages. Indeed, the individual needs are badly adapted to the offer of many content suppliers, who want to reach a large audience and thus offer contents that are sometimes very

heterogeneous. Moreover, studies show that the majority of the users regularly revisit a small number of Web sites (around 5 to 10) compared with the total of visited sites. This was partially taken into account by some content suppliers who proposed personalized accesses (MyAltavista, MyCNN, etc). The idea of aggregation involves that of collecting some parts of dispersed Web sites and of associating them on a single access. The main difficulty is splitting Web pages into semantically homogeneous parts and describing them well. These descriptions will make it possible to sort contents or to guide the process of re-assembly. These various stages can be more or less automated by systems, based on artificial intelligence using XML standard. In the same field, we can also note the image collage initiative (See collage machine Web site and Zawinski (2003)).

Another way of taking part in implicit co-operation is direct exploitation of Web contents and user activities. Project FET (Edmonds 1999), for example, proposes to study and exploit the phenomenon of collective intelligence on the Web on the same basis as operation in neural networks (associative links). The users' observed variables are taken into account (thematic profiles, reciprocity of connections, time of consultations, etc). Analysis of this information can be used to make recommendations to users or to drive the automated-organization of the network or to evaluate informational interaction on the Web.

F. Heylighen (1999) also explores in his paper the phenomena of collective intelligence and its computation in the Web. The paper proposes the construction of a collective mental Map containing the current state of problems to solve possible actions and preferred ones. It calculates, in particular, a co-occurrence matrix of links contained in Web pages consulted by the users, as well as their browsing path's analysis. All this information is used to produce collaborative filters that propose new relevant Web sites to the users or to feed an optimized research agent. L. Terveen et al (1998) propose web site co-referencing as an indication of emerging collaboration (density of the associated graph). They showed for example that co-referencing was very limited in a commercial context. (See also K Nakata et al (1998)).

7.3 Combining Information Selectivity and Access Speed

The exploitation of proxy cache contents to optimize information selectivity was already studied (Dodge 1995, Bradford 1999, Frakes et al 1992, Lancieri 1997, Wittenburg 1995), but these studies mainly deal with indexing of textual contents. The objective is to supply, as in the case of a search engine, the URL of documents (document reference and not the document itself) contained in the cache from keyword inputs by users. Even if, by chance, the referenced document is still in the cache, there is no external process allowing its life span to be known, foreseen or modified. This is an important difference with regard to our system where objects are taken from the cache and deposited in a local web server, having more controlled management rules and being adapted to the user needs. Finally, a project similar to our approach is the Mandala system (J. Helfman 1999). People identify groups of images visually and share them with Mandala by dragging them between windows. Groups of image representations are stored as image-maps, making it easy to save visual bookmarks, site indexes, and session histories. Mandala also uses proxy caches to optimize selectivity and access speed to information. Even if the concerns are similar, the main difference could be that our system focuses more on low-level human computer interaction (implicit aspect) whereas Mandala focuses on high level (explicit End user interactions).

We have also the Peer-to-Peer (P2P) cooperative model that allows selectivity and access speed to be combined, depending on content popularity and cooperation. The basic principle of direct exchange of resources between user computers on the network is not a new one. Indeed, the initial principle of the Internet network is based on the same philosophy. However, its evolution has transformed the web into a very big central system since main Web accesses are made to a limited number of servers (portals, university, etc.). Favoured by the recent high-speed connections and powerful computers, P2P has returned to the first Web vocation where every computer is client and server. Popularized by the free initiative (Napster, Gnutella, etc.), this subject has drawn a lot of manufacturers (Sun JXTA project, IBM

P2P Telework environment, etc, see Newstrove 2003). One of the problems of P2P is its distributed management (all users) with all the possible consistency, security, ethical and legal drawbacks for a company. The solution is not evident because a more homogeneous management would tend to normalize and make interactions more rigid, which would go against the P2P philosophy. The major advantage of this type of collaborative model is that every one can take advantage of each other's activities without modifying his working habits. The absence of the middleman facilitates the information exchange and favours its spontaneity. The peer-to-peer type of cooperation seems to be explicit since access to information results from an explicit human choice. In fact, the global availability of information over the peer-to-peer network is much more difficult to model. It not only depends on one direct user action since the more the information is replicated over the P2P network the faster it can be downloaded. Therefore, the level of information availability results from its popularity level that mostly depends on implicit cooperation.

8 Conclusion

In a few words we can say that our main contributions are the following: inspired by the natural processes of memorisation and forgetting, we have presented an approach that deals with the complexity of human-human interaction mediated by computer technologies. We have shown that strong links exist between selectivity, access speed to information and locality in an environment originally designed to be distributed and global. We have studied the interest and methods of reusing weakly structured information as well as the potential of implicit cooperation and collective intelligence. This approach replaces a part of computer algorithmic complexity with natural implicit human activity. The resulting symbiotic mechanism saves processing capacity and provides enhanced services.

More specifically, this paper has described a new device mainly based on the association of two basic network components: a proxy cache and a Web server. The existing proxy cache of an organization can be more efficiently used in order to catalyze and reuse users' collective intelligence. This reuse strategy allows large and pertinent multimedia objects to be delivered, objects that are otherwise expensive to obtain (time to seek, to download, and resources). The results often integrate objects contained on recent sites, which have not yet been indexed by regular search engines. Moreover, by recycling already downloaded objects this system contributes to a necessary effort of reducing Web bottlenecks for the benefit of the majority.

The advantage of this system is to help adequately contextualized information (i.e. Pre-knowledge) to be easy accessible. To some extent, it is comparable to reducing distances between people. Less effort is necessary to exchange, less time to wait, and less time to seek. So time wise, the trial becomes free since errors of appreciation on content cost less. If search and download of content last too long, people are discouraged and do not look at it. Such techniques that spare time render the media less present and facilitate spontaneity. If we push this theory to the extremes, media would be so powerful that it would tend to make us forget its actual presence and allow instantaneity as in face-to-face contact. But, contrary to the telephone or videoconferencing, interaction takes place asynchronously between people. Thanks to externalisation of their memory, people can communicate with each other not only if they are far away but also if they are not present at all (spatiotemporal ubiquity).

In the informational domain, just like in the physical world, recycling and reusing is the strongest tool against pollution and waste of resources like energy and time. In the informational domain reusing can only be achieved by correct memorisation, which requires an adequate memory management. Finally, the active mirror is an example of how a purely mechanical system combined with human action (Internet document, knowledge generation) can expand this human action. Instead of being manually selected, organised and presented, knowledge is automatically managed thanks to the powerful combination of

implicit human action and a mechanical engine. This is probably possible because a human being has a mechanical/physical part that can easily be interfaced with computers.

With regard to legal and ethical concerns, we saw that there are technical solutions but also that these solutions have their limits. But more globally, one of the best means of preserving human rights, including privacy seems to provide better information access. Knowing the risks, the potential solutions and the possible answers to aggressions is a form of protection. It is well known that the first action of a tyrannical government is to control the spread of information. Instead of using a unique centralised information management our system directly uses the anonymous human capacity as management power. Even if it is always possible to make improvements, we think that the active mirror is a reasonable approach combining enhanced services for user and privacy preservation.

Another aspect is the re-aggregation of contents that offers a new view of knowledge facilitating creativity. Claude Burguelin (2001) explains, for example, how text can be unconsciously used by readers as heuristic to reconstruct their memory or to expand its boundaries. Even if our system made an approximate reconstruction and presentation of user contextualized multimedia contents, it can help make associations and suggest ideas.

We have also introduced metrics (output of cooperation) that provide a better view of mechanisms underlying implicit cooperation. From a purely social point of view it is not always easy to measure human interaction and to validate assumptions related to access to knowledge behaviour. From this point of view our approach can be seen as a help to social investigation.

As we see it, beyond the technical and economical aspects, this study also has implications for legal or social domains.

9 References

Akamai Corporation Web Site; http://www.akamai.com

Bestravos, Azer (1996); *Middleware support for Data-Mining and knowledge discovery in large scale distributed systems*, In proceedings of ACM SGMOG' 96 Data-Mining Workshop.

Bonabeau, Eric / Dorigo, Marco and Theraulaz, Guy (2000) *Inspiration for optimization from social insect behaviour*. Nature, Vol. 406, juillet 2000, pp. 39-42

Bradford, Clare / Marshall, Ian W (1999) *A bandwidth friendly search engine*, Proceedings IEEE international conference multimedia computing and systems (ICMCS99) Florence Italy

Brickley, Dan (2000) W3C Collaboration, Knowledge Representation and Automatability research group Cooper, Ian / Nottingham, http://www.w3c.org/Collaboration/overview.html

Mark (2001) Wrec IETF working group http://www.wrec.org

Burgelin, Claude (2001) *Comment la Littérature Réinvente la Mémoire (how literature reivente the memory)*, Article from the Science magazine La Recherche Special Edition *'La Mémoire et l'Oubli' (Memory and forgetting),* July August 2001

Cabin, Pierre (1998) *The communication, state of the art*; Collective book directed by; Human science editions

Caching. Com Web site (2002) Internet Caching resource center http://www.caching.com

CacheNow Campaign Web site http://vancouver-webpages.com/CacheNow/

CacheFlow corporation Web site http://www.cacheflow.com/technology/

Collage Machine web site: http://www.csdl.tamu.edu/ecology/combinFormation/about.html

Cormack, Andrew (1996) Web Caching, A Report, University of Wales, Cardiff http://www.jisc.ac.uk/index.cfm?name=acn_caching

Danzig, Peter B (1995) *Massively Replicating Services in Wide Area Inter network*; Technical report university of South California (USC)

Davison, Brian D (1999) *A Survey of Proxy Cache Evaluation Techniques*, Proceedings of the 4th International Web Caching Workshop

De Rosnay, Joel (1995) *L'homme symbiotique (the symbiotic man)* Ed. du Seuil 1995

Dodge, Chris / Marx, Beate and Pfeiffenberger, Hans (1995) *Web Cataloguing through cache exploitation and steps towards consistency maintenance*; Computer Networks and ISDN systems 27

Edelman,Gerald M. (2000) *The Biology of Consciousness (Biologie de la Conscience)*. Editions Odile Jacob,1992 Re-edited in 2000.

Edmonds, Bruce / Gruendlinger, Leor / Heylighen, Francis (1999); *Supporting Collective Intelligence on the Web: design, implementation and test of a self-organizing collaborative knowledge system; Evolutionary Collaborative Knowledge Project for FET Open Proposal*; http://www.cpm.mmu.ac.uk/~bruce/bsi/

Fielding, Roy T / Gettys, James / Mogul, Jeffrey C / Frystyk Nielsen, Henrik / Masinter, Larry / Leach, Paul J / Berners-Lee, Tim (1999) Hypertext Transfer Protocol -- HTTP/1.1, Request for Comments 2616;

Frakes, William B. and Baeza-Yates, Ricardo (1992) *Information Retrieval, Data Structure and Algorithms,* editors, Prentice Hall

Fuchs, Christian (2003) *Co-operation and Self-Organization* TripleC international Journal (vol 1, n 1) 2003; http://triplec.uti.at/articles/tripleC1(1)_Fuchs.pdf

Gettys, Jim / Berners-Lee, Tim / Frystyk Nielsen / Henrik (1999) *The Web Consortium Problem Statement on Propagation, Replication and Caching*; http://www.w3.org/Propagation/Activity.html

Gwertzman, James (1995); *Autonomous replication*; Senior Thesis Harvard university; also available at http://www.eecs.harvard.edu/~vino/web/push.cache/

Helfman, Jonathan (1999) *Image Representations for Access and Similarity-Based Organization of Web Information*, also available at http://hci.ucsd.edu/lab/hcipapers/JoH1999-1-dis.pdf

Heylighen, Francis / Joslyn, Cliff A / Turchin, Valentin F (2002) Principia Cybernetica Web site http://pespmc1.vub.ac.be/ http://www.w3.org/Protocols/rfc2616/rfc2616.html

Heylighen, Francis (1999); *Collective Intelligence and its Implementation on the Web: Algorithms to Develop a Collective Mental Map*. Computational & Mathematical Organization Theory. Vol. 5, no. 3

Heylighen, Francis (1996) *The World-Wide Web as a Super-Brain: from metaphor to model*, Cybernetics and Systems '96

Himmelsbach, Guy (1999) Ariane search engine Web Site http://www.ariane6.com/

Hu, Wen-Chen / Chen, Yining / Schmalz, Mark S / Ritter, Gerhard X (2001) *An overview of the World Wide Web search technologies*, In the proceedings of 5 th World Multi-conference on System, Cybernetics and Information, SCI2001.

Kelly, Terence / Chan Yee Man / Jamin, Sugih and MacKie-Mason JK (1999) *Biased Replacement Policies for Web Caches: Differential Quality-of-Service and Aggregate User Value,* University of Michigan, fourth International Web Caching Workshop, San Diego, California.

Lancieri, Luigi (1997) *Interactive shared bookmark*; Proceedings of international conference WebNet97, Toronto

Lancieri, Luigi (2000) *Memory and forgetting, two complementary mechanisms to characterize the various actors of the Internet and their interactions*; PhD Thesis, university of Caen

Lancieri, Luigi (2001) *The concept of informational ecology or the interest of information reuse in the company*. In the proceeding of the third International Conference on Enterprise Information System (IEEE, AAAI) Portugal

Lancieri, Luigi / Berthier Bonnel, Nicolas / Stumme, Ludovic (2001 b) *To exploit the collective intelligence thanks to the Co-operative replication*; In proceedings of International Conferences on Info-tech & Info-net ICII2001-Beijing.

Lawrence, Steve / Gilles, C. Lee (1999) *Accessibility and distribution of information on the Web*; in *Nature*, Vol. 400, pp. 107-109, see also http://www.wwwmetrics.com.

Lévy, Pierre (1994) *L'intelligence collective, pour une anthropologie du cyberspace*, La Découverte, 1994. Translated from the French by Riikka Stewen in http://www.hnet.uci.edu/mposter/syllabi/readings/levy.html

Levy, Pierre Links on the practice and theory of collective intelligence, by Pierre Lévy, http://mikro.org/Events/OS/wos2/Levy-pp/liensIC.html

Moreover corporation Web Site http://w.moreover.com/

Network Appliance Corporation Web Site; http://www.netapp.com

NewsHub corporation Web Site http://newshub.com/

Net2one corporation Web Site http://net2one.com/

Nakata, Keiichi / Voss, Angi / Juhnke, Marcus and Kreifelts, Thomas (1998) *Collaborative Concept Extraction from Documents*, In proceedings of the 2nd Conference on practical aspects of knowledge management (PAKM98).

Neal, Donald (1996) *The Harvest Object Cache in New Zealand* Computer Networks and ISDN Systems, volume 28, p 1415 - 1430,

P2P Newstrove Web site http://p2p.newstrove.com/

Rochat, Philippe / Thompson, Stuart (1999) *Proxy Caching based on object location considering semantic usage,* in proceedings of Web caching workshop

Ronkainen, Pirjo (1998) *Attribute Similarity and Event Sequence Similarity in Data Mining*, Technical Report C-1998-42, University of Helsinki

Saillard, Nicolas (2003) *Optimization of architectures of replications by the characterization of the traffic*, PhD Thesis, university of Caen

Schacter, Daniel.L. (1996) *Searching for memory: The brain, the mind, and the past*, New York: Basic Books.

Simon, Herbert Alexander (1983) *Reason in Human Affairs*, Stanford University Press

Sperber, Dan (1996) *Explaining culture: A naturalistic approach*. (Oxford: Blackwell). French version: La contagion des idées (Paris, Odile Jacob)

Snippets corporation Web Site http://www.snippets.com/

Squid project Web Site http://www.squid-cache.org/

Sullivan, Danny / Sherman, Chris (2004) *Survey on search engine* http://searchenginewatch.com/

Tewari, Renu / Dahlin, Michael / Vin, Harrick M / Kay Jonathan S (1998); *Beyond hierarchies: Design consideration for distributed caching on Internet*; Austin Univerity of Texas technical report; TR98-04

Terveen, Loren G / Hill, Will C (1998) *Evaluating Emergent Collaboration on the Web*, in Proceedings of CSCW'98 (Seattle WA, November 1998), ACM Press.

Touch, Joe (1999) *LSAM project Web Site*; http://www.isi.edu/lsam

Wei, Tang (1999) *Search engine survey*, http://www.cc.gatech.edu/~wtang/research/papers/search_eng.pdf:

Wittenburg, Kent / Das, Duco / Hill, Will / Sead, Larry (1995) *Group Asynchronous Browsing on the World Wide Web*, in proceedings of the 4 th International W3C Conferences Boston USA, see also http://www.w3.org/Conferences/WWW4/Papers/98/

Wolpert, David / Tumer Kagan (1999) *An Introduction to Collective Intelligence*, Tech Report NASA-ARC-IC-99-63. Also Available at http://ic-www.arc.nasa.gov/ic/projects/collective-intelligence.html

Yodlee Corporation Web Site http://www.yodlee.com

Zaïane Omar R (1999) *Resource and knowledge discovery from the Internet and multimedia repositories*, PhD Thesis, Simon Fraser University, Canada

Zawinski, Jamie (2003) *Web Collage Web site*: http://www.jwz.org/webcollage/